Title: Using Think-Alouds to Create a Better Measure of Biology Reasoning

Dai, T., Van Boekel, M., Cromley, J. Nelson, F. & Fechter, T., (2018). Using think-alouds to create a better measure of biology reasoning. *SAGE Research Methods Cases*, January 12, 2018. doi: 10.4135/9781526437167

Ting Dai
University of Illinois at Chicago

Martin Van Boekel
Search Institute

Jennifer Cromley
University of Illinois at Urbana-Champaign

Frank Nelson
Temple University

Tia Fechter
Pacific Metrics Corporation

Author's Note
The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160335 to University of Illinois, Urbana-Champaign. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

SAGE Research Methods Cases Psychology Submission for Consideration

Case Title

Using Think-Alouds to Create a Better Measure of Biology Reasoning

Author Name(s)

Ting Dai [a], Martin Van Boekel [b], Jennifer Cromley [c], Frank Nelson [d], Tia Fechter [e]

Author Affiliation & Country of Affiliation

[a] University of Illinois at Chicago, [b] Search Institute, [c] University of Illinois at Urbana-Champaign, [d] Temple University, [e] Pacific Metrics Corporation

Lead Author Email Address

Email: tdai@uic.edu

Discipline: D3

Sub-discipline

Educational Psychology [SD-Psych-6]

Academic Level

Postgraduate

Contributor Biographies

Ting Dai is an assistant professor in the Department of Epidemiology and Biostatistics at Temple University. Her research, at the intersection of educational psychology, measurement and methodology, and the learning sciences, focuses on changes in motivation and achievement in science and math, cognitive processes in math and science learning, and educational measurement and quantitative methodology. She also brings over 9 years of experience in educational and psychological measurement, multivariate statistics, and latent variable modeling.

Martin Van Boekel is a post-doctoral fellow in the Department of Educational Psychology at the University of Illinois at Urbana-Champaign. His research interests lie at the intersection of cognitive, educational, and social psychology. Throughout his PhD his research focused on two distinct and overarching goals. The first goal is to understand how people learn when correct new information conflicts with incorrect prior knowledge. The second goal of his research is to investigate the factors that support positive youth development.

Jennifer Cromley is an associate professor in the Department of Educational Psychology at the University of Illinois at Urbana-Champaign. Her research program explores the role of both cognition and motivation in science achievement and retention. In her work, she investigates the role of these variables in STEM achievement, its relation to retention in STEM majors, and cognitive and motivational interventions to improve student outcomes.

Frank Nelson is an associate professor in the Department of Biology at Temple University. His research program explores the cognitive processes that underlie science learning. In his work, he investigates students' understanding of complex science concepts . Overall, his work is at the intersection of educational psychology, cognitive science, and biology.

Tia Fechter is Director of Psychometric Services and Test Assembly at Pacific Metrics Corporation – an ACT Assessment Technologies company. Her research interests include item response theory test equating methods, item parameter drift detection, generalizability theory applications, and validity studies.

Published Articles

Not Applicable.

Abstract

This case study describes how our research team conducted the qualitative think-aloud approach (or cognitive pretesting) to obtaining cognitive validity evidence for a biology inference-making and reasoning measure for undergraduate students. The main goal of our work was to gather high-quality student think-aloud data of reasoning while they were completing the biology measure to complement quantitative psychometric data of the measure. We provide a detailed case study that demonstrates the development, refinement, and application of our think-aloud protocol. This case study also reveals the challenges we encountered, and discusses the valuable lessons we learned from solving various problems in this process. With this case study, we aim to provide student researchers with not only a brief introduction to basic theories of think-aloud research, a procedural description of how to conduct a think-aloud study for measurement validation purposes, but also a thoughtful discussion of challenges and possible solutions to assist in think-aloud research design.

Learning Outcomes

By the end of this case students should be able to

-- have a clear understanding about how to conduct a think-aloud approach (or cognitive pretesting) to obtaining cognitive validity evidence for an educational or psychology measure,

-- conduct a qualitative think-aloud study for measure validation purposes, and

-- modify the provided example to fit the specifics of their own think-aloud research.

Case Study

Project Overview and Context

This case study describes our mixed methods approach to gathering validity evidence for a biology inference-making and reasoning measure (IMRB). We focus on describing our collection of qualitative evidence using a think-aloud ("Say everything you are thinking while

you answer the question") approach, to complement the quantitative validity evidence that is easy to find in the literature. This study employed purposive sampling, involved multiple sites, and used a novel method for gathering validity evidence – this approach has both benefits and challenges, which we will describe below.

For STEM (Science, Technology, Engineering, and Mathematics) majors, being able to engage in reasoning and inference-making are critical skills for the academic success and retention in a science major (e.g., Dai & Cromley, 2014; Cromley, Perez, Wills, Tanaka, McNamara Horvat, & Tancredi-Brice Agbenyega, 2013). However, there are few biology inference measures for use with undergraduate STEM students. Popular measures of scientific reasoning tend to focus on aspects of the scientific method and rely on some core biological content knowledge. While these are important aspects of becoming a biologist, the emphasis of introductory biology courses is *reasoning with biological principles.* The inference-making and reasoning in biology measure (IMRB) is *specifically rooted in reasoning about and with biological principles*, but does not require prior knowledge of specific biological terms.

Like the creation of any new measure it is critical to provide sufficient validity evidence, so that we, and others using this measure, can be confident in using the tool that it measures what we propose. Common quantitative sources of this evidence factor analysis, IRT, and DIF. These quantitative approaches allow us to examine the latent dimensionality of a measure, calculate item difficulty and discrimination, and evaluate fairness of the items.We have completed all of these analyses on the IMRB, and from a psychometric perspective this measure behaves as we would hope and expect. However, relying exclusively on quantitative evidence can be problematic. For example, Muis, Duffy, Trevors, Ranellucci, and Foy (2014) used a think-aloud to gather validity evidence for the Discipline-Focused Epistemological Beliefs Questionnaire, a popular self-report measure of epistemic beliefs. The results from this think-aloud found that participants were not interpreting and responding to the items as

researchers using this measure had intended. These misaligned interpretations critically affect the inferences that can be drawn from the data, and could not have been identified by quantitative methods alone. Therefore, we used a think-aloud protocol to provide evidence to support reliability, validity, and other measurement properties; and to determine designated cognitive validity of items and obtain evidence for predictive validity of the IMRB. The think-alouds will reveal the students' thinking processes, information they did not understand or are not bringing to bear, jumping to conclusions, choosing an item that was literally stated in the text (rather than drawing an inference in an inferential question), as well as misunderstandings or misinterpretations of test instructions or what a question is asking.

Research Context

This research took place over one academic year at two large universities. One university is located in the Midwest, the other on the East Coast. Participants were students who took an introductory biology course 1 to 3 years before their participation in the study. The think-alouds were conducted in a quiet environment, and in a one-on-one setting lead by a member of the research team.

Research Design

Our research used a mixed methods approach. The IMRB was created in 2006. Since its creation it has been used as a pre and post test measure in a variety of interventions (e.g., Cromley et al., 2016), giving us a large set of data (N = 2,274) to explore the quantitative aspect of the validity evidence. Differential item functioning, or DIF, is a concern for many measures. The possibility that group membership (e.g., ethnicity/race, sex, first-generation-college-student status) influences performance on an item instead of ability alone would be very problematic for the IMRB. In order to ensure that we had a study sample that could be used to shed light on possible sources of DIF we engaged in purposive sampling when recruiting participants for the think-aloud study. Using this approach we recruited a total of

80 students (40 females) with prior experience in taking introductory biology courses to complete the think-aloud. Within these 80 students, we recruited 20 each of African-American, Asian, white, and students of "Other" races within the female and male groups (e.g., 10 African-American females and 10 African-American males, etc.). Our sample also consisted of 40 each of first-generation (neither parent with a Bachelor's degree) and non-first-generation (at least one parent with a Bachelor's degree) college students.

Each think-aloud participant completed the IMRB on a desktop computer while engaging in the think-aloud procedure. The participant was instructed to answer 15 multiple-choice questions on the IMRB at their own pace, and was provided with a pencil and scrap paper as needed. Many participants were dubious of the think-aloud process at first, but after engaging in a practice problem on a subject matter not related to biology, were able to focus on the task, and were able to think aloud with few prompts while completing the IMRB. After completing the IMRB, participants completed a short demographic questionnaire. The participants were allotted one hour to complete the entire task, but were allowed more time if needed. Most participants were able to finish in approximately 40 minutes. In the section that follows we will focus on the think-aloud process because even though the think-aloud is recommended for survey and measure design as part of the piloting phase of tool creation, it is rarely used as a means to gather strong validity evidence.

Conducting the Think-Aloud

There is a long tradition of using think-alouds in psychology research (Willis, 2015). One of the foundational texts on the think-aloud protocol is *Protocol analysis: Verbal reports as data* by Ericsson and Simon (1993). We drew heavily from the examples provided in this text when designing our think-aloud protocol. It is important to think about the various ways think-aloud data can be collected before engaging in your study. First, data can be collected in real time, as the participant is taking the test. In this approach, the interviewer administers

the session and prompts the participant as necessary to ensure the participant reads aloud and verbalizes what she or he is thinking during the process of completing the questionnaire. The prompts typically exclude pronouns such as “me” and “us” to ensure minimal social interaction between the interviewer and the participant involved in the process (example prompts can be found below). Second, the participant can complete the test as they would in a normal testing situation, and then only once they have completed, return to the test and recount what they were thinking while completing the test. This retrospective prompting and recount approach is researcher-structured rather than spontaneously generated by the participant, and involves social interaction. Or third, they could have a combination of both, where participants think aloud as they complete the test, then once finished the researcher could further probe the participant to expand on their thoughts, and explain unexpected behaviors. This combined approach involves both active report by the participant and reactive response to researcher’s questions, and involves social interaction in the retrospective part of the study.

After several rounds of piloting the procedure, we decided to use the first option, participants engaged in think-aloud while completing the test. We selected this option because the pilot test takers seemed to be more focused, explained more of their thoughts, to a large extent worked at their own pace without excessive interruptions, approached the think-aloud in a more natural way with minimum influences from the researcher, and appreciated the fact that we were being considerate about their time. In the very rare occurrence that participants did something unusual, i.e. ignoring the diagrams, or selecting an option without reading the entire question or answer options, the researcher asked the participant to explain what they were thinking retrospectively when they did this.

The IMRB Measure

The IMRB is a 15-item measure used to assess reasoning and inference-making in the biology domain (an example item is provided below). Participants are provided a short segment of text, or a diagram and asked questions related to that material. Because the material required to answer the question is provided within the stimuli associated with the questions, background knowledge is not required. Even if the student does guess what the biological term means (e.g., lymphocyte), she or he can still reason through the answer to the question.

Antigens Interact With Specific Lymphocytes, Inducing Immune Responses And Immunological Memory.

Although it encounters a large repertoire of B cells and T cells in the body, a microorganism interacts only with lymphocytes bearing receptors specific for its various antigenic molecules. The "selection" of a lymphocyte by one of the microbe's antigens activates the lymphocyte, stimulating it to divide and to differentiate. Eventually the lymphocyte forms two clones of cells. One clone consists of a large number of effector cells, short-lived cells that combat the same antigen. The other clone consists of memory cells, long-lived cells bearing receptors specific for the same antigen. This antigen-driven cloning of lymphocytes is called clonal selection.

Each lymphocyte helps fight _____.

a. a group of pathogens (ex: fungi).

b. B cells and T cells only.

c. cloning which can damage the immune system.

d. only one specific type of pathogen or foreign molecule.

Obtaining the Sample

We had very a specific population in mind for this study, therefore we engaged in purposive sampling. Because participants were required to have relatively recent experience with the introductory biology course (within 3 years of participation in this study) we relied heavily on the support of past biology professors to help us recruit participants. We found some participant groups especially difficult to fill. In those instances, we asked professors and course coordinators for specific suggestions so we could send individualized recruitment emails. We also asked participants to advertise our study to friends they think might be interested.

The Think-Aloud Procedure

Our research was being conducted across two sites, therefore it was important that the testing be as consistent as possible across the two locations. This means that a clear, and easy to follow protocol had to be created. Several researchers associated with this study had extensive experience with the think-aloud process before beginning this study. Even with their combined expertise in this area, many rounds of piloting and revising were required before an easy to follow protocol was finalized.

Our complete protocol can be found below. In order to highlight the parts that can be easily overlooked while conducting this study we have embedded ***Notes for Researchers*** (bolded and italicized in brackets) in the protocol. Immediately following the complete protocol, we discuss in detail the variety of lessons learned from conducting this think-aloud study.

--PROCEDURE BEGINS--

<u>Setting up before participant arrives:</u>

Materials:

1. Audio recorder
2. Clipboard with protocol and prompt options

3. Pen and paper for general note taking: two sets (for both researcher and participant- have date and participant number on participant's form)
4. Consent form
5. Computer and Internet set up so participant can respond to IMRB and demographic form in Blackboard
6. Researcher's chair positioned as far back from participant chair as possible (to discourage conversation) but close enough to see what participant is doing/writing

[NOTE: Both the researcher and the participant may find this room setup to be a little strange, but unlike ordinary social interactions, the goal here is to gather the participant's thoughts while responding to the questions, not to encourage conversation.]

Things to check before participant arrives:

A. Check to see that the audio recorder has batteries.

B. Check to see that there are enough copies of the study materials for the current participant and those who are schedule for future studies.

C. Register participant in Blackboard using their NetID.

D. Open and then minimize IMRB Measure_Interview Version document (used during interview phase if needed)

E. Prepare a hard copy of the IMRB Measure in case of computer failure

<u>When the participant arrives: (expected time 60 minutes)</u>

1. Welcome participant, introduce yourself, and walk the participant through the consent form. Take a photo of the consent form and email or text photo to the participant. Have students log into the Blackboard website.

2. Explain the procedures to the participant:

“You are being presented with multiple-choice questions based on material taught in IB150. We are interested in learning about how students think about and answer questions such as these, in order to improve the measure we have created. I want you to answer the questions as if you were being tested on them in a course such as IB150. There is no time limit for answering the questions, but we expect it may take you about 25 minutes to answer them.

In this experiment I am interested in what you are thinking as you answer questions. In order to do this I will ask you to THINK ALOUD as you are reading and answering the questions. What I mean by think aloud is that I want you to say out loud EVERYTHING that you are thinking from the time you start reading each question until you give an answer. I would like you to talk out loud CONSTANTLY from the time I present each question until you have given your final answer to the question. Say everything that goes through your mind, even if you think it seems irrelevant. I don’t want you to try to plan out what you say or try to explain your thoughts, but verbalize your thoughts as they occur. Just act as if you are alone in the room speaking to yourself. It is important that you keep talking. If you are silent for any length of time I will remind you to keep talking aloud. Also, if I can see you doing something, but you are not verbalizing what you are doing, I will ask you to say what you are doing. Do you understand what I want you to do?” [***NOTE: Answer any questions as needed.***] “For this phase, I can’t answer any questions about the questions or help you with them, although we very much appreciate your feedback on how the questions and answer options are worded. So if you have a question, I want you to ask it. I will respond saying “Sorry, I cannot help you with that”, but if you think of a question, it is important to verbalize it. Please remember that it is very important to say everything you are thinking while you are working on this task. To get you used to the process of thinking aloud, I have a sample question here. Before you

begin, I would like to turn on the recorder so that I can capture everything that you are saying."

[NOTE: It is important that the researcher accurately gives all instructions above WITHOUT reading off the script. Make eye contact to ensure the participant responds to and understands the instructions.]

3. Turn on audio recorder.

"Today is [DATE] at [TIME] o'clock. I am working with [PARTICIPANT NUMBER]."

[NOTE: It is important that the participant's name not be used or recorded so that we can ensure anonymity.]

"You can begin when you are ready. When you have answered the question click on the arrow on the right side of the screen to progress to the next question"

SAMPLE.

The purpose of inferential statistics is to take the results of a test on a sample and make an inference to a population.

AND

A sample must be drawn from a specific population.

THEREFORE

a. Sampling is done so that inferences can be made to a population.

b. The results of inferential statistics allow for generalization to the population from which the sample was drawn.

c. The results of inferential statistics allow for generalization to the whole population.

d. Inferential statistics only allow us to make inferences about the sample we drew.

[NOTE: It is very important that a good practice question is used. If the question is too easy the participant may produce the answer automatically which would not yield an opportunity to practice thinking aloud. It is possible and beneficial to use multiple practice

questions. We wanted the participants to be able to complete the entire experiment in under an hour, therefore we opted for a single example that utilized a similar structure as items that could be found on the IMRB only testing a different subject domain.]

If a participant is silent while working through the question:

Use a mixed interval prompt period (i.e., once after 4 seconds, another after 2 seconds, and so on). If the participant is silent prompt them with the following question: "What are you thinking right now?" or "Say what you are thinking." If the participant can be seen doing something related to the test, say "Say what you are doing" in order to get a clear impression of their strategies.

At this point in the study it is very important not to use other prompts, the participant's verbalizations should not be led by us in anyway. DO NOT ASK PARTICIPANTS TO "EXPLAIN" OR "TELL YOU" WHAT THEY ARE DOING.

[NOTE: The prompts outlined above are prompts commonly used in other think-aloud studies. These prompts all functioned well during the piloting phase. However, when working with participants we noticed that the prompt "What are you thinking right now?" often resulted in the participant immediately stating which option they were thinking about. Because we do not want the phrasing of our prompts to influence the participants' think-aloud we decided to stop using this prompt on all subsequent think-alouds. Therefore, based on our piloting experiences we relied exclusively on the two prompts: "Say what you are thinking." and "Say what you are doing." Using an irregular schedule of prompting ensures the participant does not get into the habit of only verbalizing at a certain frequency.]

4. "Great. Now I would like you to do the same thing for several more questions. I will present you the items, one at a time. Please remember to say everything that comes to your

mind as you read the question and as you are working out your solution. Here is a piece of paper and pencil just in case you need it. Do you have any questions before we begin?"

Present the participant with the first item. *During this portion of the task, we will be asking the participants to engage in a think-aloud. This means that we want participants to think aloud while they are reading the item, and while they are answering the question.*

Use a mixed interval prompt period (i.e., once after 4 seconds, another after 2 seconds, and so on). If the participant is silent prompt them with the following question: "What are you thinking right now?" **[*NOTE: As previously mentioned, we discontinued the use of this prompt based on its unintended influence on the participant's think-aloud process*]** or "Say what you are thinking." If the participant can be seen doing something related to the test, say "Say what you are doing" in order to get a clear impression of their strategies.

5. As the participant is answering the items, note behaviors on a given item in order to get a sense of strategies participants are using as they are taking the test. For example, look for things like: pausing on an option for a length of time; immediately eliminating a few options; rereading the question several times; etc. We want to understand how students are responding to our items, if they are interpreting them correctly, so their behaviors may mark parts of an item/items that need to be revised. Some students will not have notes because they are verbalizing all of their thoughts. These notes will largely be physical things participants are doing that cannot be captured by the think-aloud. Remember to use the "Say what you are doing" prompt in these situations to encourage the participant to verbalize their processes.

6. Once participant finishes with entire test, and if you have taken notes about their behaviors say:

"Now that you have answered all the questions, I want to ask you some questions about your thinking while you were answering them. Because we are trying to understand your own thinking process, there is no right or wrong answer to the questions I'm going to ask."

Maximize the IMRB Measure_Interview Version document and scroll to the items corresponding to your notes.

If no notes were taken, turn off the recorder and proceed to the demographics form.

7. *After the participant has completed the test:*

Say the following:

"Now that you have answered all the questions, I want to complete a quick questionnaire."

Navigate the participant to the demographic form and let them complete the questionnaire.

8. Ask participant if they have any questions.

9. Participants will receive their amazon gift via email. Make sure that participants have provided their NetID. Let participants know that they should receive the email from the research team regarding their Amazon gift card, and that this email will request a confirmation.

--END OF PROCEDURE--

Lessons Learned

Need for Clear Protocols

This may seem like an obvious point, and in many ways it is. But this is especially true when engaging in a think-aloud study, particularly one that involves multiple sites. Participants often came into the study openly expressing that they have no idea what a think-aloud entails, and were worried about being able to think aloud while completing a test. We knew we needed to mitigate these feelings through our instructions and our practice question. We knew that participants may not be completely confident and experienced with this process after only one practice question, so clearly articulating the prompts that could be used by the researchers helped ensure that every participant was getting the same instructional guidance when completing the task.

It was not enough just to have a protocol written. The protocol needed to be piloted and revised multiple times. When using the protocol we also needed to be open to learning from our participants' experiences, and be open to fine-tuning the protocol. The best example of this fine-tuning comes from the identification of one of the prompts as potentially leading (discussed above).

It was a good practice for all researchers who would lead the think-aloud sessions to go over the pilot session protocol (and recording) together to learn from the piloting researcher's experience and collectively discuss issues encountered and make critical revisions to the protocol. This was particularly important for our study to maintain procedure consistency between different sites. Our pilot phase helped us identify many of the changes that needed to be made before we commenced the study. Had we had to make significant changes to the protocol during the study, it would have been likely that we would have to make those changes and restart the study.

The revised written protocol also needs to be delivered in a skillful way. It was crucial that the researcher integrates the instructions in a naturalistic way in her conversation with the participant prior to the think-aloud, rather than reading it off the script which may lead to failure in engaging the participant. The researcher should also provide plenty of opportunities for the participant to ask questions about the procedure so that participants' confusion can be properly addressed by the researcher. The timing for doing so needs to be before and after the participant does the practice question. After the practice question, it was important to provide feedback on how the participant had carried out the think-aloud with the practice question and to advise on possible improvements, e.g., speaking louder.

Difficulty in Purposive Sampling: The Importance of Relationship Building

In order to complete this study we needed a very specific sample. We would never had been able to meet our sampling requirements if it were not for the help of the introductory biology professors and the participants who advertised our study to their friends. Forging these relationships before you get to the point of data collection is key. Across both of the sites represented in this paper, researchers had long standing relationships with the biology professors. The biology professors' help in this study was completely voluntary, and they were not associated with this study in anyway. Before our request for support in finding participants, we discussed the project with them, and outlined a timeline for the study. Therefore, when we reached out and asked for support, they were not surprised by the request and were willing to help.

When we faced difficulties reaching our sampling requirements in a demographic category, our long standing relationships with others involved with the students' educational experience became very important. With their support we were able to reach out to students directly through individualized recruitment emails. Even though these were not always successful, in the end we were able to obtain our proposed sample.

Think About the Participant's Experience

When we engage in research there are many questions that we *could* or *want* to have answered. However, it is very important to consider your participants and their experience with your experiment. Even if students are being compensated for their time, we as researchers must remember that these students have other activities going on in their life. In our case, the process of engaging in a think-aloud is often new to participants, and can be pretty tiring. We wanted our participants to be focused on engaging in a high quality think aloud, while trying to do their best on the IMRB. Therefore, we needed to find the appropriate balance between providing time to learn about the process through instructions, time to practice, and still have time to complete the experiment.

Completing a test is hard enough, now imagine completing a test while saying everything that you are thinking out loud in front of a stranger who is silently sitting nearby taking notes about your actions. As a researcher that is thinking about using a think-aloud protocol, it is also very important to consider how your participant may feel throughout the experiment. If your participant feels vulnerable or exposed, this may influence what they are willing to say out loud. It is your job as the researcher to make your participants feel as comfortable as possible. Remember that research is an exchange. The participants are completing your study, giving you much needed data, and in turn you may be providing compensation of some sort, in our case a gift card. As part of this exchange, it also helps if you can highlight how your participants are contributing to our knowledge on the subject under investigation. We did this by outlining in our consent form the goals of the study, and how we will use the information they provide us with.

Be Mindful of What Comes Next

As researchers, we have had the importance of collecting validity evidence to support instruments we plan to use drilled into us over years of study. It is not uncommon to find a

host of psychometric and quantitative validity evidence sources published in journal articles. It is far less common to see a large scale qualitative look at how participants are responding to, and interpreting the items as we are doing with the IMRB. We wanted to get a complete look at our measurement tool, so that we can be absolutely confident in our measure as both a tool for research, and potentially classroom use. But this is not an easy process.

In this paper we have presented the methodological aspect of using a think-aloud. What we have not talked about is the even larger and more time consuming aspect of this work - the analysis. Once the data has been collected, it must be transcribed. This requires yet another protocol to be developed, transcribers trained, and quality control performed. Then comes the coding. The transcriptions must be parsed into thought units, a coding scheme needs to be developed or selected, coders must demonstrate inter-rater reliability, and only then can the actual coding begin. It is possible that during this phase that you realize that your coding scheme is not sufficient, and it needs to be revised. Then the whole process must start over again. Only once all of this is done can the data analysis begin. This may seem daunting, but in the end it is worth it, because you have rigorously collected validity evidence for your measurement tool, that may help you make significant improvements if need be.

Conclusion and Relevance for the Larger Field of Psychology

Using a multi-phase, mixed method approach to evaluate and modify the IMRB allows us to gather the validity evidence needed to ensure that we have created a high quality measure. The quantitative approach, IRT and DIF, allowed us to identify *which* items are troublesome, and the qualitative approach, think-alouds, allowed us to better understand the cognitive processes underlying students' thinking. This provided the opportunity to identify *why* items are troublesome and why others are high quality, why some items are easy for examinees to answer correctly and why some are harder. Taken together, these approaches significantly contribute to the validity argument supporting the utility of the IMRB, and will contribute to

our knowledge about the factors associated with high quality inference and reasoning items in biology.

Exercises and Discussion Questions

1. Why might educational and psychological researchers be interested in using think-aloud protocols (or cognitive pretesting) for measurement development and refinement?

2. What are some useful approaches to developing and revising a think-aloud protocol at the beginning phase?

3. Why is it necessary to reduce research-participant social interactions during a naturalistic think-aloud?

4. What are some challenges researchers might encounter when conducting a think-aloud study for measurement validation purposes?

Further Readings

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.

References

Cromley, J. G., Kaplan, A., Perez, T., Dai, T., Mara, K., Davidson, Y., Balsai, M. J., & Richards, E., (2016). *Improving undergraduate biology learning and achievement with a combined cognitive-motivational intervention.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 8-12, 2016.

Cromley, J. G., Perez, T., Wills, T. W., Tanaka, J. C., McNamara Horvat, E., Tancredi-Brice Agbenyega, E. (2013). Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors. *Contemporary Educational Psychology, 38*(3), 247-258.

Dai, T., & Cromley, J. G. (2014). Changes in implicit theories of ability in biology and dropout from STEM majors: A latent growth curve approach. *Contemporary Educational Psychology, 39*(3), 233-247.

Ericsson, A., & Simon, H. (1993). *Protocol analysis: Verbal reports as data (revised edition).* Cambridge, MA: Bradford books/MIT Press.

Muis, K., Duffy, M., Trevors, G., Ranellucci, J., & Foy, M. (2014). What were they thinking? Using cognitive interviewing to examine the validity of self-reported epistemic beliefs. *International Education Research, 2*(1), 17-32.

Willis, G. B. (2015). *Analysis of the cognitive interview in questionnaire design*. Oxford University Press.